AMENDMENTS TO THE CLAIMS

Claims 1, 15, 30, 33 and 37 are amended herein.

Claims 2, 11-13, 16, 25-27, 31 and 34 are canceled.

Claims 1, 3-10, 14-15, 17-24, 28-30, 32-33, and 35-37 are now pending. All pending claims are produced below.

1. (Currently Amended) A system for identifying language attributes through probabilistic analysis, comprising:

a storage system adapted to store a set of language classes, which each identify a language and a character set encoding, and further adapted to store a plurality of training documents;

an attribute modeler adapted to train an attribute model by evaluating occurrences of one or more document properties within the training documents and, for each language class, calculating a probability for the document properties set conditioned on the occurrence of the language class, the trained attribute model stored in the storage; [[and]]

a text modeler adapted to train a text model by evaluating byte occurrences within the training documents and, for each language class, calculating a probability for the byte occurrences conditioned on the occurrence of the language class, the trained text model stored in the storage[[.]]; and

a training engine adapted to calculate an overall probability for ones of the set of language classes by evaluating the probability for the document properties set based on the attribute model and the probability for the byte occurrences based on the text model.

2. (Canceled)

3. (Previously Presented) A system according to Claim 1, further comprising:

an assignment module adapted to assign the overall probability for a language class in accordance with the formula:

$$\arg\max_{cls} P(text \mid cls) \cdot P(props \mid cls) \cdot P(cls)$$

where *cls* is the language class, *text* is the byte occurrences set, *props* are the document properties, and *P*(*text* | *cls*) is the probability for the byte occurrences, and *P*(*props* | *cls*) is the probability for the document properties set.

4. (Original) A system according to Claim 1, wherein the document properties comprise at least one of top level domain, HTTP content character set encoding and language header parameters, and HTML content character set encoding and language metatags.

5. (Previously Presented) A system according to Claim 4, further comprising:

an assignment module adapted to assign the probability for the document properties set based on the attribute model in accordance with the formula:

$$P(tld, enc \mid cls) \cdot P(cls)$$

where *tld* is the top level domain, *enc* is the character set encoding and *cls* is the language class.

6. (Previously Presented) A system according to Claim 1, further comprising:

a counting module adapted to count byte co-occurrences within a training document, and determine the probability for the byte occurrences based on the byte co-occurrences.

7. (Previously Presented) A system according to Claim 6, wherein the byte co-occurrences comprise a set of trigrams, further comprising:

a probability module adapted to calculate a probability of a trigram as the number of occurrences of the trigram divided by the total number of trigram occurrences in the training documents for a language class.

8. (Previously Presented) A system according to Claim 7, further comprising:

an assignment module adapted to assign the probability for the byte occurrences set based on the text model in accordance with the formula:

$P(text \mid cls)$

where *text* is the set of trigrams and *cls* is the language class.

9. (Previously Presented) A system according to Claim 1, further comprising:

a training engine adapted to perform iterative training by providing the probability for the document properties set and the probability for the byte occurrences set respectively to the evaluation of byte occurrences and assignment of the set of language classes.

10. (Previously Presented) A system according to Claim 1, further comprising:

a back off module adapted to evaluate less frequently occurring document properties by calculating a probability for a less frequently occurring document property conditioned on the occurrence of the language class.

11. (Canceled)

12. (Canceled)

13. (Canceled)

14. (Previously Presented) A system according to Claim 1, wherein at least one training document comprises one of a Web page and a news message.

15. (Currently Amended) A method for identifying language attributes through probabilistic analysis, comprising:

defining a set of language classes, which each identify a language and a character set encoding, and a plurality of training documents;

evaluating occurrences of one or more document properties within the training documents and, for each language class, calculating a probability for the document properties set conditioned on the occurrence of the language class by an attribute model; [[and]]

evaluating byte occurrences within the training documents and, for each language class, calculating a probability for the byte occurrences conditioned on the occurrence of the language class by a text model[[.]]; and

calculating an overall probability for ones of the set of language classes by evaluating the probability for the document properties set by the attribute model and the probability for the byte occurrences by the text model.

16. (Canceled)

17. (Previously Presented) A method according to Claim 15, further comprising:

assigning the overall probability for a language class in accordance with the formula:

$$\arg\max_{cls} P(text \mid cls) \cdot P(props \mid cls) \cdot P(cls)$$

where *cls* is the language class, *text* is the byte occurrences set, *props* are the document properties, and $P(text \mid cls)$ is the probability for the byte occurrences, and $P(props \mid cls)$ is the probability for the document properties set.

18. (Original) A method according to Claim 15, wherein the document properties comprise at least one of top level domain, HTTP content character set encoding and language header parameters, and HTML content character set encoding and language metatags.

19. (Previously Presented) A method according to Claim 18, further comprising:

assigning the probability for the document properties set based on the attribute model in accordance with the formula:

$$P(tld, enc \mid cls) \cdot P(cls)$$

where *tld* is the top level domain, *enc* is the character set encoding and *cls* is the language class.

20. (Previously Presented) A method according to Claim 15, further comprising:

counting byte co-occurrences within a training document; and

determining the probability for the byte occurrences based on the byte co-occurrences.

21. (Previously Presented) A method according to Claim 20, wherein the byte co-occurrences comprise a set of trigrams, further comprising:

calculating a probability of a trigram as the number of occurrences of the trigram divided by the total number of trigram occurrences in the training documents for a language class.

22. (Previously Presented) A method according to Claim 21, further comprising:

assigning the probability for the byte occurrences set based on the text model in accordance with the formula:

$P(text \mid cls)$

where *text* is the set of trigrams and *cls* is the language class.

23. (Original) A method according to Claim 15, further comprising:

performing iterative training by providing the probability for the document properties set and the probability for the byte occurrences set respectively to the evaluation of byte occurrences and assignment of the set of language classes.

24. (Previously Presented) A method according to Claim 15, further comprising:

evaluating less frequently occurring document properties by calculating a probability for a less frequently occurring document property conditioned on the occurrence of the language class.

25. (Canceled)

26. (Canceled).

27. (Canceled)

28. (Previously Presented) A method according to Claim 15, wherein at least one training document comprises one of a Web page and a news message.

29. (Original) A computer-readable storage medium holding code for performing the method according to Claim 15.

30. (Currently Amended) A system for identifying documents by language using probabilistic analysis of language attributes, comprising:

a set of language classes, each language class comprising a language name and a character set encoding name;

a training corpora comprising a plurality of training documents;

an attribute modeler adapted to train an attribute model by evaluating a top level domain and character set encoding associated with the training documents and, for each language class, calculating a probability for each such top level domain and character set encoding conditioned on the occurrence of the each language class; [[and]]

a text modeler adapted to train a text model by evaluating co-occurrences of a plurality of bytes within the training documents and, for each language class, calculating a probability for the byte co-occurrences conditioned on the occurrence of the each language class[[.]]; and

a training engine adapted to calculate an overall probability for ones of the set of language classes by evaluating the probability for the top level domain and character set encoding based on the attribute model and the probability for the byte occurrences based on the text model.

31. (Canceled)

32. (Previously Presented) A system according to Claim 31, further comprising:

a plurality of unlabeled documents; and

a classifier classifying one or more unlabeled documents by at least one language class, comprising:

an attribute evaluator determining document properties within the documents and initializing language class probability to each document from the attribute model;

a text evaluator evaluating byte occurrences in the documents and updating the language class probability of the each document from the text model;

a pruner pruning at least one language class falling below a predetermined probability threshold; and

an assignment module assigning at least one language class based on the language class probability of each document.

33. (Currently Amended) A method for identifying documents by language using probabilistic analysis of language attributes, comprising:

defining a set of language classes, each language class comprising a language name and a character set encoding name;

assembling a training corpora comprising a plurality of training documents;

training an attribute model by evaluating a top level domain and character set encoding associated with each training document and, for each language class, calculating a probability for each such top level domain and character set encoding conditioned on the occurrence of the each language class; [[and]]

training a text model by evaluating co-occurrences of a plurality of bytes within each training document and, for each language class, calculating a probability for the byte co-occurrences conditioned on the occurrence of the each language class[[.]]; and

calculating an overall probability for ones of the set of language classes by evaluating the probability for the top level domain and character set encoding based on

the attribute model and the probability for the byte occurrences based on the text model.

34. (Canceled)

35. (Previously Presented) A method according to Claim 33, further comprising:

accessing a plurality of unlabeled documents; and

classifying one or more unlabeled documents by at least one language class, comprising:

determining document properties within the documents and initializing language class probability to each document from the attribute model;

evaluating byte occurrences in each document and updating the language class probability of the document from the text model;

pruning at least one language class failing below a predetermined probability threshold; and

assigning at least one language class based on the language class probability of the document.

36. (Original) A computer-readable storage medium holding code for performing the method according to Claim 30.

37. (Currently Amended) An apparatus for identifying documents by language using probabilistic analysis of language attributes, comprising:

means for defining a set of language classes, each language class comprising a language name and a character set encoding name;

means for training an attribute model by assigning at least one top level domain and character set encoding pairing to at least one language class for each of a plurality of training documents and calculating a probability for each such top level domain and character set encoding pairing conditioned on the occurrence of the assigned language class; [[and]]

means for training a text model by evaluating co-occurrences of a plurality of bytes within each training document and, for each language class, calculating a probability for the byte co-occurrences conditioned on the occurrence of the language class based on the attribute model[[.]]; and

means for calculating an overall probability for ones of the set of language classes by evaluating the probability for the top level domain and character set encoding based on the attribute model and the probability for the byte occurrences based on the text model.